Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 22 No. 6 (2026) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v22i06.61535

SPECIAL FOCUS PAPER

An Overview and Methodical Strategy to Counteract the Medical Data Shortage for AI Applications

Firman Menne[1], Arya Kumar[2], Devi Debyani[3] (✉)

[1]Bosowa University, Makassar, Indonesia

[2]Kalinga Institute of Industrial Technology (KIIT) Deemed to be University, Bhubaneswar, Odisha, India

[3]Sri Sri University, Cuttack, Odisha, India

devi.d@srisriuniversity.edu.in

ABSTRACT

Artificial intelligence (AI) has the power to improve healthcare systems. Additionally, AI has the potential to improve the accuracy and fairness of medical facilities. The amount of data we currently have is inadequate despite an increase in new data. There can be an issue in this area since some health-related diseases occur less frequently than others. The amount and complexity of health-related data limit our ability to collect this type of information since this collection is expensive and complicated. At times, there are either not enough subjects included in the study or, collectively across studies, there is a lack of subject numbers. As AI and health care evolve, the performance of an ML (machine learning) model will always be poor if it does not have sufficient/adequate amounts of data upon which to learn. As a result, the models may be biased and perform poorly in medical settings. This study looks at the problems caused by a lack of information in the healthcare sector. It talks about what this means and how we can fix these issues. The study also looks at how specialists in machine learning (ML) are addressing these issues and how these concepts can be used in medical settings and in the health sector through models of machine learning. This review aims to provide researchers looking to create trustworthy predictive models using ML for healthcare purposes with a useful resource.

KEYWORDS

artificial intelligence (AI), healthcare systems, medical data shortage, machine learning (ML) models, bias and accuracy

1 INTRODUCTION

The clinical application of artificial intelligence (AI) technology, especially deep learning (DL), has attracted considerable attention from the health sciences community over the past few years. This interest has been sustained despite the considerable regulatory hurdles that this field faces. These hurdles raise ethical issues such as fairness, autonomy, privacy, transparency, safety, cybersecurity, trust, and accountability. According to Santosh and Gaur [1], the interest in this field offers to provide

Menne, F., Kumar, A., Debyani, D. (2026). An Overview and Methodical Strategy to Counteract the Medical Data Shortage for AI Applications. *International Journal of Online and Biomedical Engineering (iJOE)*, 22(6), pp. 139–155. https://doi.org/10.3991/ijoe.v22i06.61535

Article submitted 2026-03-14. Revision uploaded 2026-04-09. Final acceptance 2026-04-09.

accurate solutions to the numerous problems faced by the healthcare system. These problems include the increasing administrative burden, the diminishing consultation times, the lack of personalisation in treatments, and the subjective nature of medical diagnoses and severity assessments. There are also challenges associated with the scalability of telemedicine to the global population. Topol points out that the lack of enough healthcare professionals worsens these problems. This hurts the quality of care while simultaneously hindering the move to a more personalised and human-centric form of medicine.

The primary objective of DL algorithms is to autonomously identify statistical patterns within data, facilitating the automation of complex tasks such as differential diagnosis, triage, primary care, early detection of disease outbreaks, disease monitoring, severity assessment, and even robotic surgery. However, the effectiveness of these algorithms relies heavily on the availability of extensive and representative datasets. While numerous research fields stand to benefit from access to large public datasets, the medical domain presents unique challenges due to concerns over patient privacy, regulatory constraints, and the inherent heterogeneity of medical data. These factors often result in fragmented and inadequate datasets that lack the diversity necessary for DL models to generalise effectively. As a result, the limitations of datasets often hinder the ability of DL research to extend beyond its initial context, posing significant challenges for wider application [2]. Furthermore, the quality of the data is equally important as its quantity, as models' performance is intricately tied to the data on which they are trained. Discrepancies between training data distributions and real-world data can lead to performance mismatches between experimental and real-world settings. Furthermore, the quality of the data is equally important as its quantity, as models' performance is intricately tied to the data on which they are trained. Discrepancies between training data distributions and real-world data can lead to performance mismatches between experimental and real-world settings. Additionally, concerns such as unintended biases and labelling errors in datasets can further compromise the practical applicability of models developed in research environments [3].

In the medical field, obtaining high-quality, representative datasets presents distinct challenges. These include:

Rare Diseases: A scarcity of data for rare diseases leads to an unbalanced distribution of the dataset. This scarcity hinders the ability of DL models to reliably identify such conditions. Although rare diseases individually may not be prevalent, the collective frequency of rare conditions could still be significant.

- **Expensive Data Annotation:** Data annotation is a time-consuming process; medical data annotation requires specialised knowledge and attention to detail. This process is expensive as a result.
- **Noisy Data and Labels:** Clinical data is often not standardised, and the annotation may not have good intra- and inter-rater agreement.
- **Limited Population Coverage:** Medical data sets often only cover the population from the cohort from which they were obtained and may not cover the entire range needed to cover real-world settings.
- **Technical Limitations:** Imaging devices and other data-gathering tools usually belong to a single manufacturer's product line and are made for perfect operating circumstances that do not usually exist in everyday clinical settings, where these models will eventually be used.
- **Ethical and Legal Issues:** Privacy-related laws restrict the collection and use of private medical information in AI research.

This paper provides an overview of common strategies for developing robust AI solutions in data-scarce settings. "Robustness" is defined as a model's ability to maintain stable, reliable performance in the face of challenges such as limited dataset size, the presence of rare groups, and noisy data and labels. The approaches discussed are particularly relevant to medicine, especially in the area of medical image analysis, which is one of the most active and promising fields for applying DL in healthcare [4]. The study focuses on broadly applicable methods and avoids discussing techniques tailored to narrow problem sets, such as methods designed for specific imaging techniques. Additionally, we explore model-centric approaches, like robust loss functions and self-supervision, alongside data-centric strategies, such as active learning and data augmentation. Rather than structuring the discussion by method or historical context, we address practical problems and provide solutions for each.

This paper contributes actionable insights to assist researchers and practitioners in developing robust AI systems for me.

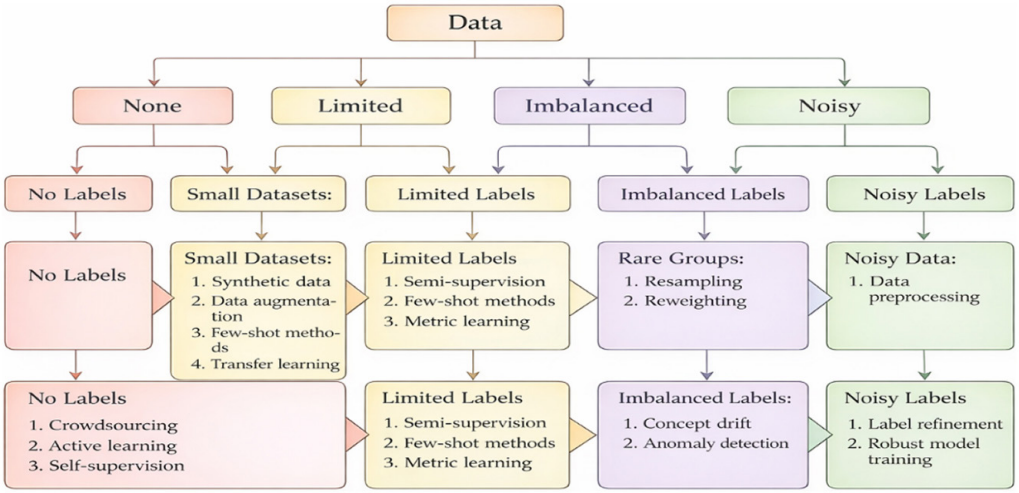

Fig. 1. Overview of data scarcity scenarios and associated methodologies

Notes: "MC" refers to model-centric, and "DC" refers to data-centric. Arrows indicate that the strategies applied in one scenario can be adapted to another.
Source: Created by authors.

2 LITERATURE REVIEW

These challenges are compounded by the limited number of available data points as the application of AI continues to grow across so many different fields and disciplines. However, reviews focusing specifically on strategies to mitigate data scarcity remain rare. One notable study explores learning-based approaches for training on small datasets [5]. While it discusses sectors particularly impacted by data limitations, such as electromagnetic and medical imaging, the review is predominantly focused on DL techniques, considering only publications post-2019 and overlooking data-centric approaches for addressing data scarcity. In contrast, data-centric approaches are discussed in a separate study by Richard et al. [6], although without specifically addressing the issue of data scarcity [7]. Meanwhile, Bansal et al. [8] concentrate on data augmentation methods as a solution to combat data scarcity. Other surveys focus on specialised scenarios, such as learning from unlabelled data [9] and few-shot learning techniques.

2.1 Systematic categorisation

This section provides a systematic framework to identify approaches to address a lack of data; it provides a decision-making framework to assist healthcare practitioners facing similar situations (see Figure 1). The propositions put forth in this section, and the framework established, could result in the classification of like methods into different categories; together, this serves as a practical resource for practitioners wishing to identify the most appropriate solutions to the specific problems they are addressing.

2.2 Scarcity of data/labels

The use case will typically dictate whether the scarcity problem relates only to the label data or applies to the entire dataset, including feature data. In some fields (e.g., medical annotations), obtaining a high-quality label often requires considerable labour (e.g., expert review and gold-standard creation (i.e., histological analysis)). In other situations, labels can be acquired through automatic methods but at a lower quality (e.g., scraping images and captions from the web). This common scenario of scarce labels justifies a separate focus, although many techniques for scarce data can also be applied to address limited labels. Section 4 covers methods for dealing with scarce data in general, while Section 5 reviews strategies specific to scarce labels.

2.3 Quantity/quality of data/labels

In addition to the type of scarcity, quantity and quality of data/labels will be critical factors affecting the ability of a model to learn from the limited or low-quality data available. For instance, in addition to extreme cases where zero data exist, there are commonly issues related to: no label (refer to Section 5.1) or a limited number of labels (refer to Section 5.2). Furthermore, the potential scarcity may also be limited only to certain subpopulations and therefore require a focus on rare populations (refer to Section 4.2) as well as imbalanced label distributions (refer to Section 5.3). There is often a trade-off between data quantity and quality, with the option of acquiring more data at the expense of lower quality. Consequently, noisy data (refer to Section 4.3) and noisy labels (refer to Section 5.4) are treated as specific forms of data scarcity.

2.4 Data-centric vs. model-centric approaches

Both architectures and training strategies, as well as data, can be improved in AI systems. Data can be improved in various ways. Two broad approaches to overcoming data scarcity are referred to as “data-centric” and “model-centric.” Data-centric approaches typically include creating additional labels via algorithms or discovering the high-value samples to learn from in the dataset. Model-Centric approaches try to enhance model performance by the use of inductive biases, including some unique form of architecture and/or modified training objectives (e.g., transfer learning). The respective terminology is indicated alongside each approach. When it comes to enhancing model robustness, Liang et al. (2020) [10] discussed the criticality of

identifying out-of-distribution data that can lead to mistakes when developing AI models that will be used in real-world settings (such as medical imaging), where the datasets used may lack the same distributions as the real-world applications they are being created for.

2.5 Training and evaluation

Once training is complete, it is essential to evaluate the model's generalisation on independent, unseen data. In the context of data scarcity, evaluation data itself may be scarce. Evaluation not only determines model performance but also the quality of performance estimates, which are crucial for making decisions. Since the quality of these estimates often lacks a control mechanism, evaluation deserves special attention. Best practices for evaluating models in situations with limited data are summarised in Section 6.

2.6 Scope limitations

This paper focuses primarily on popular strategies for building robust AI systems with limited data in the medical domain. The goal of this review is to provide practical ways to improve performance on data-limited medical applications. For this reason, the emphasis will be on well-validated and widely deployable strategies. The purpose of this document will focus mainly on medical imaging applications and will not include those areas relevant to the healthcare field, such as natural language processing.

Although some recent innovations are mentioned, highly specialised methods for specific applications are excluded. Privacy preservation, which is closely tied to data scarcity in medical contexts, is also a significant issue but is beyond the scope of this paper. For further exploration, the reader is directed to dedicated works on the subject.

3 MATERIALS AND METHODS

This section addresses situations where there are insufficient or inadequate qualitative and representative samples for traditional supervised learning, as in the case of exceptionally rare diseases.

3.1 Small datasets

The first focus is on methods that directly tackle the challenge of small datasets—those with a limited number of samples. Strategies to increase the size of such datasets include data augmentation and synthetic data generation. The quality of commonly used dermatology image datasets (e.g., Fitzpatrick17k, DermaMNIST) was studied by Abhishek et al. [11] to show how difficult it can be to establish reliable datasets for training AI models that are accurate in the context of medical imaging. The combination of a hybrid attention mechanism and learnable thresholding

developed in Ghali et al. [12] represents a significant advancement in the ability to accurately detect 3D brain tumours in the context of medical imaging.

Data augmentation (Data-Centric). Data augmentation refers to an approach used to avoid overfitting by producing variations of existing data. This method helps models to be trained on multiple training examples that have been slightly altered from their original nature, including improved performance, generalisation, and robustness. Perturbations of original data are often very similar to the type of variation that would occur in the real world, such as equipment changes, lower-quality images, and factors such as sunlight and dust exposure.

Data augmentation methods can be divided into transformation and generation techniques. Transformation techniques, discussed here, include geometric transformations (e.g., scaling, rotation, and flipping), colour space changes, random erasing, and random cropping [9]. More advanced methods, such as Mix-up, which interpolates data samples to create new ones. A second example given regarding automating the detection of incorrectly labelled data is FastDup [13], which was designed to find duplicates within a dataset. By finding duplicates, the quality of the data obtained will improve significantly.

In the medical field, data augmentation can easily be performed on common types of image acquisition data (e.g., photographs) using already existing augmentation methods. When it comes to methods of acquiring images in different ways, such as by using X-rays, EEGs, CT scans, and MRIs, however, methods specifically designed for the unique characteristics of each data acquisition method must be utilised. For example, augmenting the original data using traditional methods will not work for EEGs but will require modality-specific adversarial methods to achieve the same level of augmented data as with other modalities. Advanced frameworks such as MONAI and TorchIO are widely used in medical applications and provide a solid starting point.

Synthetic data (Data-Centric). Artificially generated information that is used to supplement or replace real-world inputs when training machine learning (ML) models is referred to as synthetic data. When there is insufficient real data available for training a model, generating synthetic data is beneficial because it allows for expanding an existing dataset without incurring additional costs for collecting more real data. The assumption is that the synthetic examples are sufficiently representative of real-world data to improve model performance. While synthetic data is sometimes viewed as a form of data augmentation, it is distinct in that it often generates entirely new data from real samples.

Generating synthetic data. A primary strategy for generating synthetic data is through the use of generative models. Throughout history, models of generating synthetic data have progressed from older types, like mixture models and hidden Markov models, to much more complex and recent forms based on DL techniques. The most advanced models used for generating synthetic data and improving real sample quality today are known as variational autoencoders (VAEs), generative adversarial networks (GANs), and diffusion models. As a result, these generative models have become incredibly useful in producing high-quality, realistic data samples that accurately resemble the characteristics of actual data, thus helping to alleviate the problem of data scarcity.

In healthcare, synthetic data has been used extensively to protect privacy while still developing AI models [14]. However, it has also proven to be an effective solution for addressing data scarcity, particularly in fields like medical imaging.

One of the key advantages of generative models is their ability to produce diverse and challenging samples that can help overcome some of the limitations associated

with traditional data augmentation techniques. Generative models come with associated problems as well; particularly, synthetic data will not necessarily mimic the distribution or visual characteristics of real-world data.

Data retrieval (Data-Centric). The term "data retrieval" describes how you can augment an existing dataset by collecting relevant information from a collection of external data sources. For example, if you currently have a small dataset with a few examples of a rare phenomenon, it may not be adequate for training/evaluating a model effectively. An approach to overcome this issue is to investigate whether there are additional examples of the same type of phenomenon in larger datasets that were created originally for other applications. These larger datasets might include publicly available datasets such as Google Image Search or those generated through social media platforms, as well as through open data repositories for research.

The idea of data retrieval is informed by techniques that are used throughout the field of image retrieval, which focuses on identifying suitable instances from large collections of images by means of query-based searches (examples can be found in the literature [15]). Historically, much of the work in the area of data retrieval has been performed using metadata as the primary filtering mechanism to create filter sets, using the attributes of the instance to develop filters (e.g., using labels and descriptive words) so as to create candidates to be retrieved [16]. In more recent times, though, the technology available for data retrieval has moved away from being primarily based on metadata to being more dependent on unsupervised data retrieval methods that utilise similarity metrics to identify candidate instances for retrieval.

In addition, the value of the newly retrieved samples may be significantly less without validation by a qualified expert, as there can be significant variability in both the relevance and the quality of the data, depending on the source.

Transfer learning (Model-Centric). Transfer learning is a very effective solution for the problem of insufficient data by allowing for the use of pre-trained models that have been trained on similar tasks. Essentially, by training a model on one task (the source), you can use that model as the starting point for another task (target), typically in a new domain or at a different level of complexity. One of the biggest advantages of using transfer learning is the savings of time, resources, and data in training a model from scratch. Transfer learning works on the basis of the belief that the features learned during the training of the source task will be of value to the performance of the model for the target task, regardless of the degree of similarity between the source and target tasks.

Domain adaptation is a well-established form of transductive transfer learning that aims at transforming a learned model from one domain to another which are not entirely different from each other. Recent advances in transfer learning have created methods for efficiently fine-tuning models by leveraging techniques such as progressive layer freezing/unfreezing and Low-Rank Adaptation (LoRA) to reduce resource requirements for computation and memory through the usage of fractionalised and reduced quantities of the original training weight set.

3.2 Rare groups

A form of data scarcity that is most critically deployed in the domain of fairness is the existence of rare groups. In medical applications, there can be a lack of representation of certain demographic groups, making it imperative that the model remains consistent across all groups. Techniques for generating synthetic data, retrieving,

resampling, and reweighting can all be applied to achieve a satisfactory level of performance from the model in the presence of rare groups.

Resampling techniques for improving data samples from rare groups to equal samples from major group data. Resampling is a technique used to adjust the amount of sample data from rare groups so that it is equal to that of the majority group. Resampling can be accomplished by oversampling the rare group or under-sampling the mass/group of samples. Oversampling methods and techniques have emerged independently from augmentation methods and synthetic data creation methods and approaches. One such technique of oversampling includes the Synthetic Minority Over-sampling Technique (SMOTE) method, which creates synthetic sample data instances by interpolating between extreme instances of the rare group and instances of the rare group that are located nearest to the extreme instances of the rare group. Under-sampling methods, where sample data from the majority group has been removed from the minority group, contain random sampling methods and cluster-based sampling rules.

Model-centric reweighting. In the training of an ML model, we can assign different weights to the records used by M based on how often they occur. This is known as the reweighting of a sample. Generally, the average weight assigned is based on an inverse frequency for a rare group. Alternatively, when the datasets themselves are reweighted or randomly changed, both of these concepts result in the creation of a dissimilar dataset when compared with the original dataset. A common instance of model reweighting would be in a classification task where the goal is to achieve those classifications across all groups. In classification tasks, this can be accomplished through resampling and/or the use of weighting factors based upon the inverse frequency of each group or their individual sample difficulty or informativeness as measured by some clearly defined metric.

3.3 Noisy data

Noisy data refers to data containing inaccuracies, inappropriate samples, or error-to-model performance; therefore, low-quality data can have a detrimental effect on the performance of the model. The principal method for dealing with noisy data is to perform data cleaning, which is the process of identifying mistakes in data and correcting them. In the context of machine learning, it is critical to perform data cleaning on datasets, especially in areas where the quality of data is critical for organisations like healthcare. Labelling errors are frequent issues that occur with large sets of medical data (such as ECG data). Doggart et al. [17] created an automated system that can greatly enhance the accuracy of these data sets by being able to identify discrepancies in labels and making adjustments to them as needed in real-time.

Data cleaning (Data-Centric). Data cleaning is the process of finding and resolving incorrect data within the dataset. Common examples of this include finding the following types of errors: irrelevant samples, close-to-duplicates, and label errors. For example, consider the case of an X-ray image being accidentally included as one of the images in a dataset that should only contain images from dermatoscopic procedures or a close-to-duplicate image being erroneously included in both a training and test dataset. These types of errors will introduce noise into the dataset and have a huge negative impact on the accuracy of the model.

There are many tools and methods available to assist with the data cleaning of a dataset. Many automated methods have been developed specifically for performing data cleanup using ML algorithms to automatically find and fix data errors, such as

identifying outliers, finding duplicate records, and correcting incorrectly labelled data. In the section about automating the detection of wrongly labelled data, Mueller et al. [18] described two new methods for automating the identification of incorrect labels within a data set. These are Cleanvision and FastDup, and they were presented along with evidence of their effectiveness in improving the quality of a dataset [14]. Therefore, determining how clean the data you are using is and why performing data cleaning is required is an important consideration, particularly in high-stakes environments, such as healthcare, where clean and accurate data are essential to reliable predictions made by the predictive models.

4 RESULTS

4.1 Scarcity of labels

Label scarcity is the topic of this section; label scarcity is a problem, especially within the medical industry, that occurs when there is a limited number of high-quality labels for validated datasets. In a case where only the machine's data is available without an associated label, the model can only make predictions based on its own internal dataset, whereas there are no labelled datasets for the models' validation, and so label scarcity is typically a challenging technical problem for a company that has raw or unlabelled data. Recently, there have been many significant advances in multimodal models, which have affected the classification of data as being either a sample or a target. However, most researchers still believe in labelling as a relatively separate activity to observing data; we will continue to refer to this method because the concepts of label scarcity are useful, and methods to prevent scan size from limiting the quality of our models will still be relevant in multimodal use cases.

4.2 No labels

Defining labels will generally be referred to as assigning labels to a given set of data as a target from which an ML algorithm will learn to predict. Manual data labelling is a time-consuming and labour-intensive task. The number of labels can be very large on very large datasets, so the accuracy of the labels will determine how well the models trained on those given datasets represent what humans can do in that domain. Therefore, achieving a balance of quality, quantity, and costs with respect to acquiring labels for ML projects becomes a difficult issue for most researchers since it will ultimately come down to familiarity with the specific field of use for the ML project.

For this reason, many researchers have created labelling methods that will improve the speed at which you can collect labels and/or reduce noise from the dataset. The label acquisition methods outlined above are efficient and can scale labelling activities for datasets, despite being unable to provide the same level of quality as multiple expert annotators or gold-standard labelling methods. In addition to improving dataset size and coverage, efficient labelling can still produce samples with less than perfect labelling quality.

Crowdsourcing (Data-Centric). Crowdsourcing is an established and widely used method for assigning labelling tasks by breaking down large-scale labelling tasks into smaller pieces and having many non-expert annotators assign labels to

those smaller pieces either in person or through the use of online platforms like Amazon MTurk and CentaurLabs. In-crowd sourcing typically has a larger portion of the work performed by in-person annotators, whereas online platforms tend to have a limited number of annotators who perform a proportionally small part of the annotating.

Multiple medical fields have been able to apply crowdsourcing successfully for creating large, labelled datasets. However, the primary disadvantage of crowdsourcing is that the labels contain lower quality than an expert would provide for the same data. To mitigate this disadvantage, most crowdsourcing applications employ quality control techniques to ensure that the label quality is as high as possible; some common quality control methods used include requiring consensus from multiple annotators or obtaining a large number of individual annotations on each sample. Crowdsourcing remains economically infeasible, usually in cases where high-quality labels are needed across numerous samples over large datasets.

Active learning (Data-Centric). The objective of using active learning is to decrease annotation costs through an iterative process of choosing samples for labelling that will provide maximum benefit for improving future model performance. Instead of labelling every sample in the dataset, only those that are predicted to improve future model performance will be selected for labelling. As a result of this focused labelling process, model training is accelerated while maintaining approximately equivalent model performance over the estimated full dataset of sample labels. Active learning can significantly accelerate the labelling process while ensuring that the selected data samples are highly informative.

The core of active learning involves selecting samples based on their informativeness, which can be determined by assessing their uncertainty or representativeness in the data distribution. The uncertainty of a sample is typically measured by examining how confident the model is in its predictions, while representativeness is based on the distribution of labelled and unlabelled data. Active learning uses these measures to ensure that the most informative and diverse samples are selected for labelling. Techniques such as clustering and traversal strategies are commonly used for this purpose.

Active learning has been widely applied in medical fields, including electronic health records, clinical text classification, and breast cancer diagnosis. However, active learning faces challenges, such as the “cold start” problem, where initial models, trained with few labelled samples, struggle to estimate sample informativeness accurately. This issue can lead to suboptimal sampling during the early stages of the training process. Techniques such as self-supervised pre-training have been proposed as an effective way to address this issue by providing a better starting point for active learning.

Self-supervision (Model-Centric). Self-supervised learning (SSL) utilises a technique whereby a model produces sample labels from an unlabelled dataset by achieving a pretext task; a constructed supervised task that, once completed, allows the model to understand general features of the data to be used later to achieve project- or debrief-specific model training.

Augmentation-driven SSL includes methods such as contrastive learning, where models are trained to recognise different augmented views of the same sample. SimCLR is an example of this, in which augmented versions of the same image are compared with other randomly sampled images to learn representations that are robust to data variation. On the other hand, prediction-driven SSL focuses on predicting missing parts of the data based on the rest. For instance, masked autoencoders are trained to predict the missing parts of the input data after masking certain portions.

In medical applications, SSL is increasingly used to improve model predictions and reduce the reliance on annotated data. SSL has been used for tasks such as differential diagnosis, segmentation of anatomical structures, and identification of medical conditions. Although SSL provides a significant advantage by minimising the need for labelled data, it is computationally intensive and requires domain-specific expertise to design effective pretext tasks. Earlier, it was stated that Susmelj et al. [13] found self-supervised learning (SSL) to be useful for extracting features from raw, unlabelled data and that this technique could be used to improve self-supervised learning for medical image analysis.

4.3 Limited labels

The strategies described in Sections 5.1 (absence of labels) and 4.1 (small datasets) can provide assistance even when only minimal quantities of labelled examples exist. Techniques designed to produce new methods for providing additional labels as described in Section 5.1 can provide additional labelled examples, while techniques designed to help with small dataset size as described in Section 4.1 can provide more examples. However, we now focus on methods specifically developed to handle situations with limited labels, such as semi-supervised learning (SSL), metric learning, and meta-learning. These methods overlap with few-shot learning (FSL), which is particularly important in medical fields where emerging conditions or rare diseases must be identified with minimal labelled data.

Semi-supervision (Data-Centric). The purpose of SSL is to utilise both labelled and unlabelled examples in order to improve performance. For example, an individual example is utilised for training, which employs a small number of strong label examples and a large number of weak label examples (i.e., unlabelled); the weak label examples are created automatically and generally will not be as strong as the strong label ones. The combined use of strong examples and weak examples should conduct a stronger learning process by allowing the combination of both to establish more generalisable results of a Logistic Regression-based classifier or other classifiers.

Within SSL are a number of assumptions. Two of these fundamental assumptions are (1) the smoothness assumption (i.e., items that are near each other in feature space are more likely to belong to the same class) and (2) the cluster assumption (i.e., items in a cluster are more likely to belong to the same class). Both of these are critical to the overall success of SSL processing; therefore, if correct, this information will guide the learning process.

Metric learning (Model-Centric). Metric learning is the process of defining a similarity measurement between example data sets and is particularly useful when there are typically not enough labelling examples present, as it gives information about how example data sets relate to each other. The main assumption of metric learning is that similar examples will have similar representations within the learned space, allowing for the better use of limited labelled data through a metric learning algorithm.

A general-purpose representation of features is learned first within a metric learning algorithm through one of the following: transfer learning, SSL, or self-supervised learning. After a general feature representation has been established, the model will learn a task-specific similarity measurement so that it can compare new sample instances to sample instances that have been previously labelled (support set). Therefore, the model will use similarities between a test sample and a support set to make predictions.

Metric learning has successfully been employed in many medical imaging applications, including: organ segmentation of CT scans; the detection of cardiac regions in MRI scans; and the diagnosis of various dermatoses from dermoscopic pictures. Research into cross-domain feature learning is aimed at improving the ability of metrical learning models to generalise across many types of medical datasets.

Meta-learning (Model Centric). Meta-learning does both optimisation of the model and algorithm together to improve performance on new tasks when limited amounts of data exist. The meta-learning paradigm is especially beneficial in cases where there is very limited data available, since it provides a method for the model to be able to effectively adapt to any new task very quickly, since there is very little prior training on that task when compared with the general case. Few-shot learning has recently emerged as a key method in the field of medical imaging; few-shot learning allows AI models to learn from a small number of labelled training examples. A systematic review by Pachetti and Colantonio [19] of a few-shot learning algorithms validates the potential usefulness of few-shot learning in performing various tasks, such as segmentation and diagnosis, on medical images.

Meta-learning organises the learning experience into two levels. The first or inner level consists of a base-learner (student) performing the task of interest, while the second or outer level consists of a meta-learner (teacher) evaluating and refining the base-learner's learning process. Meta-learning has been applied very successfully to medical imaging. A primary disadvantage of meta-learning methodologies is that episodic training is very computation-intensive and complex. However, there is evidence to suggest that simple MAML meta-learning baselines using well-trained models will still yield substantial improvement. Thus, there may be substantial benefits to utilising meta-learning in data-scarce settings in some medical settings.

4.4 Imbalanced labels

When the target variable in ML features is under-represented, the scarcity of certain data subsets can cause significant issues. This situation lies at the intersection of limited labels (refer to Section 5.2) and rare groups (refer to Section 4.2) and is referred to as label imbalance, also known as class imbalance in classification tasks, or the long tail of conditions in medical diagnoses. The methods discussed in Sections 4.2 and 5.2 can be applied, including those referenced therein. Common causes of label imbalance in the medical field include rare conditions that produce long-tailed distributions and demographic factors, which can lead to an over-representation of more common conditions. For instance, focusing on a particular region might result in the neglect of more diverse medical conditions.

Concept hierarchies (Data-Centric). The granularity of concepts in a classification task can contribute to label imbalances. The number of available data samples for finer distinctions is often much smaller. For example, there are fewer cases of narcolepsy compared to sleep-wake disorders in general. Organising labels in a concept hierarchy and leveraging this structure can significantly improve ML model performance. A system that highlights the relationships among classes is often beneficial for both ML models and human experts.

Medical ontologies are frequently used in clinical diagnostics and healthcare. Consequently, many of the methods using hierarchies of concepts find their greatest application in medicine and biology when applied to electronic health records. It is easy for less complex concept hierarchies to be embedded within an AI application. Fortunately, numerous open-source and commercial options exist to aid practitioners

with this process, although individual use cases still require a very large number of resources to adapt these options. Basic methods for integrating knowledge graphs into an ML framework have been demonstrated to produce large benefits.

Novelty detection (Model-Centric). In cases of highly imbalanced labels and long-tail distributions, where certain samples cannot be successfully classified, it becomes essential to recognise samples outside the model's scope. Novelty detection methods can help identify data points that fall beyond the model's original context. This mechanism is valuable in many practical applications as a safety check before a model provides outputs that influence important decisions. Novelty detection does not require detailed information about potential deviations; it only needs to learn the characteristics of the original dataset.

We distinguish between novelty detection, which can operate with a dataset containing no novel samples, and anomaly detection, which only works with data that includes known anomalies. Related concepts, such as open-set recognition and out-of-distribution detection, also rely on classification outcomes. Novelty detection methods can be categorised into four types: density-based, distance-based, reconstruction-based, and classification-based. Density estimation methods model the data distribution and infer novelty by identifying low-density data points. Distance-based approaches use distance metrics to detect novel samples. Reconstruction-based methods rely on reconstruction errors to identify novel instances. Finally, classification-based approaches treat the original dataset as a single class, using boundaries to identify deviations.

4.5 Noisy labels

Labels are thought to be noisy when they have inaccuracies or errors within them. Numerous reasons can lead to labels having errors, for example, human errors in annotating data, inherent ambiguities in an object's label, or the way the label was created. An example of this is in the medical field. In the medical area of interest, researchers find low agreement when annotating the image, and there are also a lot of labelling tasks that are very difficult to do correctly [20]. Label noise negatively affects the model's capability to produce good results; labels with noise will lead to reduced quality of the model, less ability to generalise, and the potential for producing misleading findings. Label cleaning and using robust loss functions are two approaches to decrease the impact of noisy labels. Both label cleaning and robust loss functions can be used to address the issue of label noise, but using label cleaning will clean up the noisy labels, and using robust loss functions will adapt the training procedure to accommodate the noisy label data.

Label cleaning (Data-Centric). Label cleaning involves identifying and correcting samples that have been assigned incorrect labels. Label errors can lead to faulty evaluations and affect the quality of the training process. Label errors are common in practice—recent evaluations of benchmark datasets in ML found an average label error rate of 3.4%. Label cleaning is a crucial subtask of data cleaning.

Error detection typically relies on comparing predicted labels to ground truth values through confusion matrices. Samples with low recognition rates or minority class samples may be removed. There are also newer approaches, such as supervised contrastive learning for label error correction and confident learning, which estimates noise using probabilistic thresholds to correct labels. Tools for label error detection have been applied in areas like dermatology and automated electrocardiogram interpretation. While manual label cleaning can improve data quality, it is

time-consuming, error-prone, and may suffer from low inter-annotator agreement. Caution is advised when applying algorithmic approaches to clean labels due to their potential biases.

Robust loss functions (Model-Centric). To improve model learning in the presence of noisy labels, an appropriate learning objective is essential to prevent data poisoning from mislabelled samples. Robust loss functions are designed to minimise the impact of label noise, with theoretical guarantees that they can recover the correct result under noisy conditions. The Cross-Entropy (CE) loss function, commonly used for classification tasks, is not robust and may lead to poor performance when labels are unreliable.

While the mean absolute error (MAE) is a basic and robust alternative to CE, it doesn't always work well with deep neural networks (DNNs). For this reason, practitioners will typically use CE variants such as generalised cross-entropy (GCE) and reverse cross-entropy (RCE). Other approaches involve combining robust and non-robust loss functions to create an active-passive loss (APL) or using likelihood-based loss functions that dynamically adapt to noisy labels. Robust loss function candidates are easy to add to an existing ML pipeline with few modifications.

5 EVALUATION

A reliable estimate of how well a model will perform in real-world use is called an evaluation. In traditional ML evaluation procedures, part of the data is designated for evaluation purposes, while the remaining portion of the data is used to build the model. This methodology works well with large data sets. However, there are problems with bias and sample stability when you have a smaller sample size.

Quality and integrity of the evaluation set are paramount when working with scarce data. Using data and label cleaning methods, or manual curation (refer to Section "Semi-supervision (Data-Centric)"), will help maximise the reliability of evaluation results. In addition, we need to ensure that there is no information leakage between the evaluation set and training set so that there are truly independent samples.

In the past years, federated learning has been a useful resource of information when developing personalised solutions (or making predictions) about healthcare. One example of this is how federated learning has been used to develop a time-sensitive federated co-predictor for predicting which children with asthma will develop complications in the future [19] [20] [21] [22]. This example demonstrates that federated learning can effectively provide personalised predictions about children's future outcomes while maintaining the confidentiality of children's individual records.

6 DISCUSSION

Data scarcity is a significant impediment to the development of AI systems in medicine because of ethical and biased distributional challenges, which require creative and innovative solutions [23]. Among the potential solutions for mitigating these challenges are active learning, synthetic data generation, and self-supervised learning. The challenge will be determining when these techniques will be appropriate for use. To address the systematic data policies that apply to all phases of the data pipeline (collection through validation).

Standardised processes, such as the MIDaR (Medical Imaging Data Readiness) Scale, are critical for improving the quality of medical imaging data. The combination of model-centric and data-centric approaches to solving data scarcity, including label-cleaning and robust loss functions, represents a potentially necessary solution. Finding ways to apply a combined approach and gain clinical expertise requires further exploration.

It is critical to have a multidisciplinary approach to addressing data scarcity in healthcare, involving physicians, ML researchers, and policymakers, to ensure that AI models are valid for clinical practice and accurately represent the real-world.

7 CONCLUSION

In the medical field, there is always a shortage of data, making it very difficult to build reliable AI systems. The purpose of this review is to describe several of these mitigation techniques based on the following problem areas: (1) not enough samples of data; (2) imbalanced sample distributions; (3) noisy data samples and labels. We have provided an overview of existing and emerging methods along with their uses, benefits, and limitations (particularly regarding medicine). This review is intended to help researchers navigate the difficult terrain of data scarcity and increase collaboration with other researchers.

8 REFERENCES

[1] K. C. Santosh and L. Gaur, *Artificial Intelligence and Machine Learning in Public Healthcare: Opportunities and Societal Impact*, Springer Nature, 2022. https://doi.org/10.1007/978-981-16-6768-8

[2] A. C. Yu, B. Mohajer, and J. Eng, "External validation of deep learning algorithms for radiologic diagnosis: A systematic review," *Radiology: Artificial. Intelligence*, vol. 4, no. 3, p. e210064, 2022. https://doi.org/10.1148/ryai.210064

[3] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "AI in health and medicine," *Nature Medicine*, vol. 28, no. 1, pp. 31–38, 2022. https://doi.org/10.1038/s41591-021-01614-0

[4] D. Zha *et al.*, "Data-centric artificial intelligence: A survey," *ACM Journals*, vol. 57, no. 5, pp. 1–42, 2025. https://doi.org/10.1145/3711118

[5] N. Bendre, H. T. Marín, and P. Najafirad, "Learning from few samples: A survey," *arXiv preprint arXiv:2007.1548*, 2020. https://doi.org/10.48550/arXiv.2007.1548

[6] M. Richard *et al.*, "Prevalence of most common skin diseases in Europe: A population-based study," *Journal of the European Academy of Dermatology and Venereology*, vol. 36, no. 7, pp. 1088–1096, 2022. https://doi.org/10.1111/jdv.18050

[7] F. Garcea, A. Serra, F. Lamberti, and L. Morra, "Data augmentation for medical imaging: A systematic literature review," *Computers in Biology and Medicine*, vol. 152, p. 106391, 2023. https://doi.org/10.1016/j.compbiomed.2022.106391

[8] A. Wang and O. Russakovsky, "Overwriting pretrained bias with finetuning data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 3957–3968. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2023/papers/Wang_Overwriting_Pretrained_Bias_with_Finetuning_Data_ICCV_2023_paper.pdf [Accessed: Mar. 3, 2026].

[9] F. Gröger *et al.*, "Towards reliable dermatology evaluation benchmarks," in *Proceedings 3rd Machine Learning for Health Symposium*: *Proceedings Machine Learning Research*, vol. 225, 2023, pp. 101–128. [Online]. Available: https://proceedings.mlr.press/v225/groger23a [Accessed: Mar. 3, 2026].

[10] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," *arXiv preprint arXiv:1706.02690*, 2020. https://doi.org/10.48550/arXiv.1706.02690

[11] K. Abhishek, A. Jain, and G. Hamarneh, "Investigating the quality of DermaMNIST and Fitzpatrick17k dermatological image datasets," *Nature Scientific Data*, vol. 12, no. 1, pp. 1–21, 2025. https://doi.org/10.1038/s41597-025-04382-5

[12] S. A. Ghali, A. El-Zaart, and L. Affara, "Hybrid attention and learnable thresholding for 3D brain tumor segmentation," *International Journal of Online and Biomedical Engineering*, vol. 22, no. 3, pp. 73–93, 2026. https://doi.org/10.3991/ijoe.v22i03.58359

[13] P. Siwek, P. Skruch, and M. Długosz, "A clustering-based data reduction for the large automotive datasets," in *27th International Conference on Methods and Models in Automation and Robotics (MMAR)*, Międzyzdroje, Poland, 2023, pp. 234–239. https://doi.org/10.1109/MMAR58394.2023.10242489

[14] F. Gröger, L. Amruthalingam, S. Lionetti, A. A. Navarini, F. Ille, and M. Pouly, "A review and systematic guide to counteracting medical data scarcity for AI applications," *Computer Methods and Programs in Biomedicine Update*, vol. 8, p. 100220, 2025. https://doi.org/10.1016/j.cmpbup.2025.100220

[15] E. Pizzi, S. D. Roy, S. N. Ravindra, P. Goyal, and M. Douze, "A self-supervised descriptor for image copy detection," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 14512–14522. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2022/papers/Pizzi_A_Self-Supervised_Descriptor_for_Image_Copy_Detection_CVPR_2022_paper.pdf [Accessed: Mar. 5, 2026].

[16] H. W. Goh, U. Tkachenko, and J. Mueller, "CROWDLAB: Supervised learning to infer consensus labels and quality scores for data with multiple annotators," *arXiv preprint arXiv:2210.06812*, 2023. https://doi.org/10.48550/arXiv.2210.06812

[17] P. Doggart, A. Kennedy, E. Foreman, D. Finlay, and R. Bond, "Automated identification of label errors in large electrocardiogram datasets," in *2022 Computing in Cardiology (CinC)*, 2022, pp. 1–4. https://doi.org/10.22489/CinC.2022.321

[18] M. Ceraudo, P. Anania, A. Prior, D. Criminelli Rossi, and G. Zona, "Modified endoscopic diving technique without the traditional irrigation system in endoscopic cranial base surgery: Technical note," *World Neurosurgery*, vol. 127, pp. 146–149, 2019. https://doi.org/10.1016/j.wneu.2019.03.276

[19] E. Pachetti and S. Colantonio, "A systematic review of few-shot learning in medical imaging," *Artificial Intelligence in Medicine*, vol. 156, p. 102949, 2024. https://doi.org/10.1016/j.artmed.2024.102949

[20] H. Wang *et al.*, "A comprehensive survey on deep active learning and its applications in medical image analysis," *Medical Image Analysis*, vol. 95, p. 103201, 2024. https://doi.org/10.1016/j.media.2024.103201

[21] P. K. Shukla, S. Jain, and S. Kalra, "DeepAsthmaNet: A time-aware federated prognostic framework for personalised paediatric asthma risk stratification in primary care," *International Journal of Online and Biomedical Engineering*, vol. 22, no. 3, pp. 114–132, 2026. https://doi.org/10.3991/ijoe.v22i03.57779

[22] G. Hacohen, A. Dekel, and D. Weinshall, "Active learning on a budget: Opposite strategies suit high and low budgets," in *39th International Conference on Machine Learning*, Baltimore, Maryland, USA, PMLR 162, 2022, p. 02794. https://doi.org/10.48550/arXiv.2202.02794

[23] A. Kumar and P. Banerjee, "Leading with intelligence: How AI is reshaping creative innovation in organizations," *Prabandhan: Indian Journal of Management*, vol. 18, no. 9, 2025. https://doi.org/10.17010/pijom/2025/v18i9/174843

9 AUTHORS

Dr. Firman Menne is with the Faculty of Economics and Business, Bosowa University, Makassar, Indonesia (E-mail: firman@universitasbosowa.ac.id).

Dr. Arya Kumar is a Sr Assistant Professor at School of Economics and Commerce, KIIT Deemed to be University, Dr. Kumar, is renowned for his interdisciplinary mastery of both theoretical and practical domains. His adeptness in project management, supported by government funding, and his exemplary research acumen have garnered numerous best paper accolades at prestigious forums. Holder of multiple patents, Dr. Kumar boasts thirteen years of pedagogic excellence, with over seventy publications in esteemed journals like Emerald, Inderscience, and Springer, indexed in Scopus, WOS, ABDC, and UGC Care. He is a certified reviewer, editorial board member, and guest editor for prominent journals. He has also authored seminal textbooks for Commerce and Management students (E-mail: arya.kumarfcm@kiit.ac.in).

Dr. Devi Debyani is an Assistant Professor at Culture and Indic Studies, Sri Sri University, specialising in English language teaching, communication skills, and cultural integration in language education. Her research examines the influence of culture on language learning, the role of modern culture in pedagogy, and the intersection of leadership and organisational behaviour. She has published in renowned journals, including European Economic Letters (ABDC), YMER (UGC Care), Shodh Prabha (UGC Care), and Folia Oeconomica Stetinensia (Scopus). Her work focuses on emotional intelligence, organisational commitment, and the impact of modern culture on language proficiency. She has presented at various national and international conferences and contributed to faculty development programmes (E-mail: devi.d@srisriuniversity.edu.in).